AFRL-IF-RS-TR-2005-201
Final Technical Report
May 2005

EPI-SPIRE: A BIO-SURVEILLANCE PROTOTYPE

IBM T.J. Watson Research Center

Sponsored by
Defense Advanced Research Projects Agency
DARPA Order No. M166

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE
ROME RESEARCH SITE
ROME, NEW YORK

STINFO FINAL REPORT

This report has been reviewed by the Air Force Research Laboratory, Information Directorate, Public Affairs Office (IFOIPA) and is releasable to the National Technical Information Service (NTIS). At NTIS it will be releasable to the general public, including foreign nations.

AFRL-IF-RS-TR-2005-201 has been reviewed and is approved for publication

APPROVED: /s/

JOHN SPINA
Project Engineer

FOR THE DIRECTOR: /s/

JAMES A. COLLINS, Acting Chief
Advanced Computing Division
Information Directorate

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	MAY 2005	Final Aug 01 – Nov 04

4. TITLE AND SUBTITLE
EPI-SPIRE: A BIO-SURVEILLANCE PROTOTYPE

5. FUNDING NUMBERS
C - F30602-01-C-0184
PE - 62301E
PR - BIOS
TA - 00
WU - 01

6. AUTHOR(S)
Murray Campbell

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
IBM T. J. Watson Research Center
1101 Kitchawan Drive, Route 134
Yorktown Heights New York 10598

8. PERFORMING ORGANIZATION REPORT NUMBER
N/A

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Defense Advanced Research Projects Agency
3701 North Fairfax Drive
Arlington Virginia 22203-1714

AFRL/IFED
26 Electronic Parkway
Rome New York 13441-4514

10. SPONSORING / MONITORING AGENCY REPORT NUMBER
AFRL-IF-RS-TR-2005-201

11. SUPPLEMENTARY NOTES

AFRL Project Engineer: John Spina/IFED/(315) 330-4032/ John.Spina@rl.af.mil

12a. DISTRIBUTION / AVAILABILITY STATEMENT
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

12b. DISTRIBUTION CODE

13. ABSTRACT ***(Maximum 200 Words)***

The objective of this project was to develop prototype technologies to detect disease outbreak resulting from bioterrorism (biosurveillance) through the analysis of non-traditional data sources. The areas of focus for IBM were:

1. Develop methodologies for evaluating the usefulness of data sources for biosurveillance
2. Identify the most promising data sources for biosurveillance
3. Investigate detection algorithms that can identify early signs of disease outbreak
4. Develop methodologies to evaluate the detection algorithms
5. Develop technologies for protecting privacy of data
6. Investigate site-based biosurveillance

We worked with Greg Glass and his team at the Johns Hopkins School of Public Health. The areas of focus for JHU were:

1. Evaluate the impact of air travel on the dispersion of communicable agents
2. Evaluate selected strategies for early identification of disease outbreaks
3. Develop methods to identify permissive environmental conditions for outbreaks of zoonotic diseases in human populations

This report will give overview coverage for all of these areas, and give pointers to the included documents that explore the areas in greater depth. The report will also include a listing of all other documentation for this project, including: PI meeting documents, site visit documents, quarterly reports, and a publication list.

14. SUBJECT TERMS
Biosurveillance, Syndromic Surveillance, bioterrorism, Non-Traditional Data Sources, Data Privacy, Site-Based Biosurveillance, Zoonotic Disease, Evaluation Methodologies, Communicable Agents

15. NUMBER OF PAGES
22

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

Table of Contents

EXECUTIVE SUMMARY

The objective of this project was to develop prototype technologies that can detect the outbreak of disease resulting from bioterrorism through the analysis of non-traditional data sources (biosurveillance). There were six specific areas of focus for the IBM team's effort:

1. Develop methodologies for evaluating the usefulness of non-traditional data sources for biosurveillance
2. Apply the above methodologies to a wide variety of data sources and identify the most promising
3. Investigate detection algorithms that can identify early signs of disease outbreak in non-traditional data sources
4. Develop methodologies to evaluate the detection algorithms with respect to timeliness and false alarms
5. Develop technologies for protecting privacy of data while retaining value for detection algorithms
6. Investigate site-based biosurveillance, i.e., monitoring a geographically-constrained site, in more detail that would be possible in, say, a city-wide context.

In addition, we worked with Greg Glass and his team at the Johns Hopkins School of Public Health. The JHU team had three main focus areas:

1. Evaluate the impact of air travel, a major source of moving large numbers of people long distances, quickly, on the dispersion of communicable agents.
2. Evaluate the utility of selected strategies to identify increases in the numbers of human cases of disease (outbreaks) more rapidly than current means provide.
3. Develop methods that could be used to identify permissive environmental conditions for outbreaks of zoonotic diseases in human populations.

This report will give overview coverage for all of these areas, and give pointers to the included documents that explore the areas in greater depth. The report will also include a listing of all other documentation for this project, including: PI meeting documents, site visit documents, quarterly reports, and a publication list.

It should be noted that this project underwent substantial revision from the original statement of work, based on direction from the program manager. In particular, we increased our effort for site-based biosurveillance and away from developing a large-scale surveillance system for city or regional contexts.

IBM FOCUS AREAS

Develop methodologies for evaluating the usefulness of non-traditional data sources for biosurveillance.

Throughout the course of the program we have investigated alternatives for evaluating non-traditional data sources and their value for early warning of bio-terrorist attacks. Our most comprehensive work was presented at the International Conference for Data Mining Workshop on Life Sciences Data Mining. We used a number of different approaches to evaluate whether the sales of Over-The-Counter (OTC) medications can be useful for early warning.

Apply the above methodologies to a wide variety of data sources and identify the most promising.

Assessing the value of particular data sources was a major goal for this program. We were able to positively evaluate sales of OTC medications, showing that it provided one week of lead time or more when compared to a gold standard data source such as physician office visits.

In additional we gave evidence that certain site-based data sources had value for biosurveillance. There are two site data sources which we consider most promising: a survey of self-assessed health and phone calls to medically-related phone numbers. Absenteeism, web queries, cafeteria sales, and traffic data, though less promising, are worthy of further study. Cough counting and utility usage appear to have less value for site surveillance.

Investigate detection algorithms that can identify early signs of disease outbreak in non-traditional data sources.

We investigated a number of outbreak detection methods for the biosurveillance application. Subsequent work explored more powerful and sophisticated approaches based on time-space clustering. We extended earlier work to use more general shapes.

Develop methodologies to evaluate the detection algorithms with respect to timeliness and false alarms.

We were one of the leaders in BIO-ALIRT in establishing approaches for algorithm evaluation, introducing the AMOC approach to the group. We also pushed for an event-based evaluation on real data for the 2003 evaluation.

Develop technologies for protecting privacy of data while retaining value for detection algorithms.

Protection of the privacy of individuals when their data is used for surveillance is of the utmost concern to our project. We have explained our approach to privacy protection, which provides guarantees about the quality of the protection while still retaining as much value as possible for the data analysis.

Investigate site-based biosurveillance, i.e., monitoring a geographically-constrained site, in more detail that would be possible in, say, a city-wide context.

We focused much of our effort on thoroughly cataloging the data sources available at sites and examining them for utility in the biosurveillance task.

JOHNS HOPKINS UNIVERSITY FOCUS AREAS

Evaluate the impact of air travel, a major source of moving large numbers of people long distances, quickly, on the dispersion of communicable agents.

We developed a capacitated-network linked, susceptible-exposed-infectious-recovered (SEIR) model. This model was calibrated successfully against previously created simulations of the 1968 pandemic of influenza. When provided with current (pre-9/11/2001) air travel usage, the global pattern of influenza dispersion was dramatically altered with significant foreshortening of the dispersion. Attempts to prevent further global spread by quarantine, following the identification of cases in a city, were likely to be unsuccessful.

This model was then applied to a hypothetical release of smallpox virus, by using the same transportation data but applying patterns of disease progression for the smallpox virus. These releases were considered both for travel patterns within the United States, as well as globally. Simulations indicate that the implementation of air travel restrictions, if done rapidly can reduce the impact of smallpox spread. However, they do not prevent its

spread to other parts of the country. Thus, if an outbreak were to occur surveillance should be immediately initiated in other metropolitan areas, as well. The extent of spread is highly dependent on the city in which the initial release occurs.

Evaluate the utility of selected strategies to identify increases in the numbers of human cases of disease (outbreaks) more rapidly than current means provide.

We examined methods incorporated both modeling and statistical analyses of reported symptoms of diseases. The capacitated network SEIR model was applied to influenza-like illness data from major metropolitan areas in the United States. These models were used to compare the predicted onset and peak of influenza in the U.S. during the 1998-1999, 1999-2000 and 2000-2001 influenza seasons with results monitored by various agencies, including the World Health Organization and the Centers for Disease Control and Prevention. For nearly all the cities examined, the model anticipated the onset and peak of the influenza season 3-10 days prior to current monitoring methods.

Spatial patterns of reported illnesses, within individual metropolitan areas, also may provide important contextual clues to the appearance of a disease outbreak, whether intentional or natural. Statistical methods to detect unusual spatial patterns were applied to health related data for several locations within the U.S. Their results were compared in a Delphi experiment, with the interpretation of expert infectious disease epidemiologists. At sites where data were abundant, these methods, such as Whittemore's T statistic identified the same number of outbreaks as the expert group but identified them earlier than the experts – suggesting these programmable methods could be beneficial strategies for automated monitoring of health data streams.

Develop methods that could be used to identify permissive environmental conditions for outbreaks of zoonotic diseases in human populations.

We developed analytical methods that could be used to determine if environmental conditions were suitable for the natural emergence of zoonotic diseases (diseases carried by wildlife and transmitted to humans). Environmental monitoring methods involved merging satellite imagery, with ground station data monitoring systems as a means to improve data quality. Time series analyses of mosquito population data sets showed that currently gathered environmental data were of sufficient quality that these populations could be accurately forecasted with the implementation of new methods of data analyses. We created cross-correlation maps for the visualization of time-lagged modeling and applied empirical Bayesian estimation models to identify spatial scaling characteristics for the analyses. These approaches were applied to identify where West Nile virus was likely to occur around the Chesapeake Bay region.

IBM/JHU Publication Summary

2002

Das A, SR Lele, GE Glass, T Shields, & JA Patz. 2002. Modeling a discrete spatial response using generalized linear mixed models: application to Lyme disease vectors. Intl J GIS 16:151-166.

Glass, GE, TL Yates, JB Fine, TM Shields, JB Kendall, AG Hope, CA Parmenter CJ Peters, TG Ksiazek, C-S Li, JA Patz & JN Mills. 2002. Satellite imagery characterizes local animal reservoir populations of Sin Nombre virus in Southwestern United States. Proc Nat Acad Sci 99:16817-16822.

Iyengar, VS 2002. Transforming data to satisfy privacy constraints. In Proceedings of the 8th ACM SIGKDD pp. 279-288.

2003

Bunnell, JE, SD Price, A Das, TM Shields & GE Glass. 2003. Geographic information systems and spatial analysis of adult *Ixodes scapularis* (Acari: Ixodidae) in the Middle Atlantic region of the U.S.A. J Med Entomol 40: 570-576.

Grais RF, JH Ellis JH, GE Glass. 2003. Forecasting the geographic spread of smallpox cases by air travel. Epidemiol & Infection 130:849-857.

Grais RF, JH Ellis, & GE Glass. 2003. Assessing the impact of airline travel on the geographic spread of pandemic influenza. European J Epidemiol 18:1065-72

Li, C-S, 2003. Epi-SPIRE: A system for environmental and public health activity monitoring. In Proceedings of ICME.

Malouin, R, P Winch, E Leontsini, G Glass, D Simon, EB Hayes & BS Schwartz. 2003. Longitudinal evaluation of an educational intervention to prevent tick bites in an area of endemic Lyme disease in Baltimore County, Maryland. Am J Epidemiol 157:1039-1051.

Shone, SM, PN Ferrao, CR Lesser, GE Glass & DE Norris. 2003. Evaluation of carbon dioxide- and 1-Octen-3-ol-baited Centers for Disease Control Fay-Prince traps to collect *Aedes albopictus*. J Am Mosq Ctrl Assoc 19:445-447.

2004

Buckeridge, D, H Burkom, M Campbell, WR Hogan, AW Moore. 2004. Algorithms for rapid outbreak detection: A research synthesis. Journal of Biomedical Informatics (in press).

Florio, EN, SR Lele, Y-C Chang, R Sterner & GE Glass. 2004. Integrating AVHRR satellite data and NOAA ground observations to predict land surface temperature: a statistical approach. Int J Remote Sensing 25: 2979 - 2994.

Ammerman, NC, Swanson, KI, Anderson, JM, Schwartz, TR, Seaberg, EC, Glass, GE & DE Norris. 2004. Spotted-Fever Group *Rickettsia* in *Dermacentor variabilis*, Maryland. Emerg Infect Dis 10: 1478-1481.

Grais RF, JH Ellis, A Kress, & GE Glass. 2004. Examining the geographic spread of epidemic influenza within the U.S. *Health Care Mgmt Sci*.7:127-34.

Fine, JB, JL Robertson, GE Glass. 2004. *Borrelia burgdorferi* exposure in asymptomatic populations of horses in Maryland and Virginia. Am J Vet Res (in press)

Henshaw, SL, F Curriero, TM Shields, GE Glass, PT Strickland, & PN Breysse. 2004. Geostatistics and GIS: Tools for characterizing environmental contamination. J Medical Syst. (in press)

Grais RF, JH Ellis, & GE Glass. The role of international air travel on the global spread of smallpox. *Int J Epidemiol*. (submitted)

Iyengar VS. 2004. On detecting space-time clusters. In Proceedings of ACM SIGKDD.

JHU Students Trained while supported by DARPA Project:
Scott Shone, PhD 2001-present. Biogeography of mosquitoes in the Chesapeake Bay region of Maryland and the emergence of West Nile virus.

Derek Armstrong, MHS 2001-2002. The ecology and epidemiology of Ebola virus infections.

Andrew Walsh, PhD 2002-present. Dynamic forecasting approaches for mosquito populations.

Rebecca Freeman, PhD Geography and Environmental Engineering. 2002. A mathematical model for the geographic spread of infectious disease via air travel.

Appendix A: EPI-SPIRE: A SYSTEM FOR ENVIRONMENTAL AND PUBLIC HEALTH ACTIVITY MONITORING

Chung-Sheng Li, Charu Aggrarwal, Murray Campbell, Yuan-Chi Chang, Gregory Glass, Vijay Iyengar, Mahesh Joshi, Ching-Yung Lin, Milind Naphade, John R. Smith, Belle Tseng, Min Wang, Kung-Lung Wu, Philip Yu*

IBM Thomas J. Watson Research Center, P O Box 704, Yorktown Heights, NY 10598
*The Johns Hopkins Bloomberg School of Public Health, 615 North Wolfe Street, Baltimore, MD 21205

ABSTRACT

Health activity monitoring (HAM) has received increasing attention due to the rapid advances of both hardware and software technologies and strong environmental and public health needs. In this paper, we describe the architecture and implementation of the Epi-SPIRE prototype, which is a novel health activity monitoring system that generates alerts from environmental, behavioral, and public health data sources. A model-based approach is used to develop disease and behavior models from multi-modal heterogeneous data sources. Furthermore, a model-based indexing technique has been developed to speed up the data access and retrieval. This system has been successfully applied to various genuine and simulated diseases outbreaks scenarios[1].

1. INTRODUCTION

Recent advances in both hardware and software technologies enable real-time or near real-time monitoring and alert generation for environmental and public health related activities. Environmental related activities include global climate change (such as global warming), deforestation, natural disaster, forest fire, and air pollution. Monitoring of disease outbreaks for public health purposes based on environmental epidemiology has been demonstrated for a number of vector-born diseases such as Hantavirus Pulmonary Syndrome (HPS), malaria, and Denge fever [1-5]. Recently, health activity monitoring (HAM) concept has also been applied to the early detection of subtle human behavior changes due to disease outbreak to provide advanced warnings before significant casualties registered from clinical sources.

The alerts generated from HAM systems are triggered through the fusion of both traditional and non-traditional multi-modal heterogeneous data sources. Traditional data includes data generated from clinical sources such as in-patient and outpatient data. Non-traditional data sources include those data collected from remote sensing (including satellite images), video/audio surveillance, and other data to enable the possibility of extrapolating human behavior.

In this paper, we describe the architecture and implementation of the Epi-SPIRE prototype, which is a novel HAM system capable of generating early warning from monitoring environmental and public health activities. A model-based approach is used to develop the disease and behavior models from multi-modal heterogeneous data sources. Furthermore, a model-based indexing technique has been developed to speed up the data access and retrieval. This system has been successfully applied to vector-born infectious disease such as HPS, pests in the agriculture area such as fire ants, and influenza. For HPS, the advanced warning for high risk regions by using a combination of satellite images and digital elevation map (DEM) can be as much as 9 months [5]. In the case of influenza, preliminary results indicate that early warnings can be generated by Epi-SPIRE using

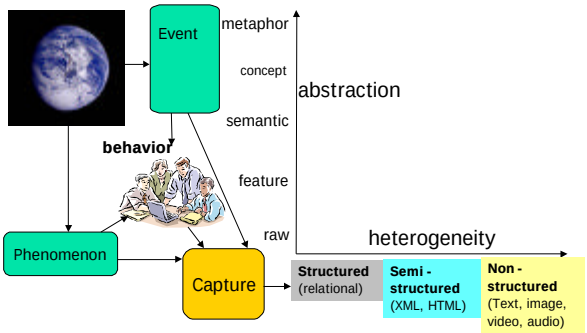

Figure 1: Process of generation of multi-modal heterogeneous data sources.

[1] This research is sponsored in part by the Defense Advanced Research Projects Agency and managed by Air Force Research Laboratory under contract F30602-01-C-0184 and NASA/IBM CAN NCC5-305.
The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied of the Defense Advanced Research Projects Agency, Air Force Research Lab, NASA, or the United States Government.

heterogeneous non-traditional data sources earlier than that can be achieved by using only traditional clinical data sources, thus demonstrating the potential benefit of such a system for public health applications.

2. PRELIMINARY ON HEALTH ACTIVITY MONITORING

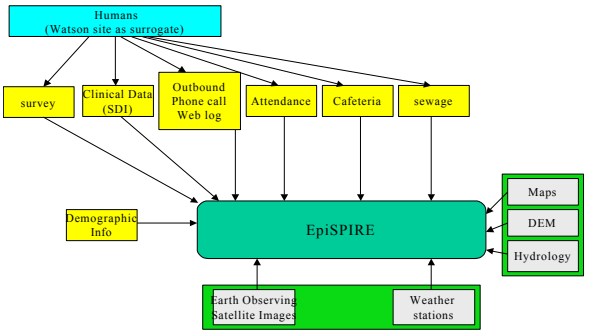

Figure 2: Epi-SPIRE environment.

The multi-modal heterogeneous data sources collected by a HAM system can come from a wide variety of sources, including (1) sensors monitoring the environment either through *in situ* or remote sensing (such as satellites) to capture the events and phenomenon as they occur; (2) data already collected for other purposes, such as e-seminar, phone records, web log, newsgroup, sewage records; (3) data collected from clinical sources such as insurance claims, in-patient and outpatient data, lab tests, and Emergency Room records.

The data sources capturing events and phenomenon related to environments and human behavior, as shown in Fig. 1, can be categorized as structured (parametric or relational), semi-structured (HTML or XML), and non-structured (text, image, audio, and video). The data can be potentially captured at various abstraction levels, including raw data (raw images or video), features extracted from the raw data (such as texture and spectral histogram from

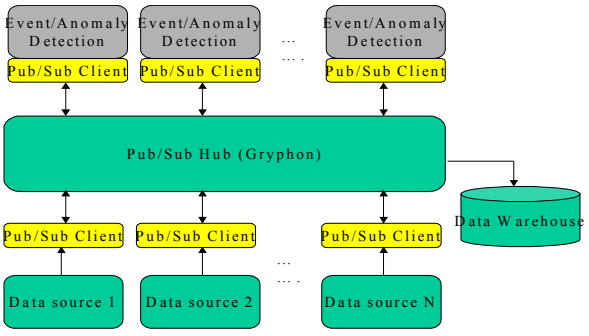

Figure 3: Epi-SPIRE architecture

satellite images), semantic (road, houses), concepts (house surrounded by bushes), and metaphors.

The main challenge in HAM is to be able to fuse multi-modal heterogeneous information sources (based on models) at different abstraction levels, generate multiple hypothesis of the models for the events, phenomenon and behaviors, and test the validity of the hypothesis using the available data. The end objective of such a system is to predict or detect an upcoming event using the model derived from the fused heterogeneous data sources.

3. ENVIRONMENTS AND ARCHITECTURE

The system environment of Epi-SPIRE is shown in Fig. 2. The Epi-SPIRE system uses (1) data collected from the natural environment (such as those collected by the satellites and weather stations), (2) data collected passively as a byproduct of human behavior (such as attendance at work or school, consumption records at cafeteria, sewage generation, web log and phone records), (3) data collected actively from probing the population that are being monitored, usually through periodic survey. In addition to the dynamic data that require real time processing, Epi-SPIRE also utilizes static data such as maps, digital elevation map, hydrology, and demographic information.

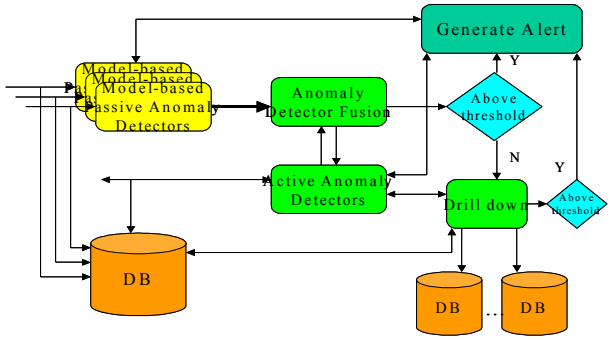

Figure 4: Alert generation process from active and passive detectors.

The system architecture for Epi-SPIRE, which is based on the use of a content-based publisher/subscriber hub - Gryphon [6], is shown in Fig. 3. All of the data sources are connected to the pub/sub hub as publisher so that the data (numeric message, text, audio, or video) from these sources can be routed through the hub to those subscribers that subscribe to these sources. All of the detectors are attached to the system as subscribers as well as publishers, so that they can subscribe to a number of data sources as well as the output from other detectors based on the topics of the data sources.

Note that each of the detectors within the system (as shown in Fig. 3) may generate alerts based on the specific charter of the detector. There is also system level alert generation that fuses the alerts generated from other detectors. The system level alert generation uses alerts

generated by both passive and active detectors, as shown in Fig. 4.

4. MODEL-BASED DATA FUSION AND DETECTION

A number of modeling techniques have been developed in this system to model the spatio-temporal risk factor to certain infectious diseases (HPS, influenza, Denge fever, and anthrax). A linear time-invariant model, $Y = a_1X_1 + a_2X_2 + \ldots + a_nX_n$, has been used to model the HPS, where each X_i represents the data itself or derived attributes/features from the multi-modal information sources, while the coefficient a_i represents the weights (relative contribution) of the attribute derived from the data. More specifically, the risk assessment model for the risk to HPS associated with a location (x,y) is:

$$R(x,y) = 0.443X_1 + 0.222X_2 + 0.153X_3 + 0.183X_4,$$

where X_1, X_2, and X_3 correspond to the pixel value of band 4, 5 and 7 of Landsat Thematic Mapper image at location (x,y), while X_4 corresponds to the elevation (in meters) from the corresponding DEM (digital elevation map). A risk map based on this model for the south western US during the summer of 1992 is shown in Fig. 5. The actual HPS outbreak took place in 1993 with more than 85% of the cases occur within those highest risk areas. In addition to the linear model, finite state machine models have been successfully developed and applied to modeling the risk to fire ants (which are harmful to both crops and livestock of the southeast US), and Bayesian network models have been developed for other infectious diseases.

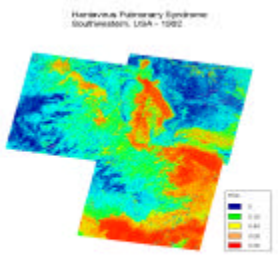

Figure 5: Risk map for Hantavirus Pulmonary

The same model for data fusion can also be used for indexing to facilitate model-based information retrieval. A model-based indexing technique, Onion [7], was developed for linear model based data fusion and retrieval and provide up to three order-of-magnitude speedups as compared to linear evaluation.

The risk map generated above provides the baseline for anomaly detection – as we are usually only interested in unexplainable anomalies. We have explored two general classes of model-based anomaly detectors (Fig. 3 and 4) that have applicability to site surveillance. The first class, which we term differential detectors, is applicable in the case where there are two or more sites that have similar behaviors. A differential detector raises an alarm when the deviation between sites becomes sufficiently large. The second class of detectors is predictive, i.e., they predict "normal" site behavior and raise an alarm if a sufficiently large deviation from normal is detected.

5. VALIDATION

Figure 6: car counting for monitoring traffic.

The Epi-SPIRE system has been validated in a genuine environment between the fall 2001 and summer of 2002 to monitor the behavioral changes of a population caused by the earliest stages of illness. Examples of such behaviors include increased absenteeism, increased inquiries for medical information, changes in eating/drinking habits, increased coughing, increased traffic for leaving the building early, and increased sewage generation. IBM T. J. Watson Research Center, which consists two sites - Yorktown and Hawthorne, and is located in Westchester County, NY (50 km north of New York City), is used in this case study. The total population for the sites is approximately 2000. All of the data collected below have been properly anonymized so that the privacy of the population being investigated is not violated.

1) A weekly survey of self-reported health level was conducted from January 2002 through May 2002, during which an email-based survey of the population was run at the Watson site. About 400 IBM employees volunteered to participate. This survey had an excellent response rate: 92% of polled employees responded the same day, 73% by noon.
2) The IBM Watson worksite requires the swiping of a badge in order to gain entry. The badge number and time of entry are recorded in a database that is maintained for security purposes. We have been receiving an anonymized version of this information since 12/2001.
3) The IBM Watson site records, for billing purposes, all phone calls made outside the site. The calling number, called number, time of call, and duration of

call are recorded in a database. A set of local medically related phone numbers was obtained from two main sources (scanned from yellow pages, internet directories). From an anonymized version of these data it is possible to count the number of calls made from Watson to medically related numbers, as well as the number of extensions that were used to place these calls.

4) The IBM Watson site records, for security purposes, all accesses to external websites at the firewall. The source IP, destination IP, and date/time of access are recorded in a database. Using an anonymized version of this database along with a manually generated list of medically related websites, it is possible to count the number of accesses to these medically related sites, as well as the number of computers from which these requests were made.
5) Consumption of cafeteria food and beverages at Hawthorne Cafeteria (one of the two sites for the IBM T. J. Watson Research Center) are recorded electronically. This cafeteria provides service to about 700 people.
6) A number of other potential data sources have been considered and undergone some preliminary evaluation. These include: site utility usage, site sewage generation, cough counting, and car counting (cars entering or leaving site). Specifically, the car counting is based on the use of the video captured from the webcam (shown in Fig. 6) in order to capture potential early departure traffic from a site. The car counter is fairly accurate except during the night or when it is raining, as shown in Fig. 7 [8].

	Total Frames	Inbound (A/M)	Outbound (A/M)	Precision
Clip 004	2417	106/95	106/100	**91.3%**
Clip 007	3856	33/32	45/44	**97.4%**
Clip 021	5460	143/143	62/65	**98.6%**
Total	11733	282/270	213/209	**96.7%**

Figure 7: Precision of car counting.

The alerts generated from these data sources are compared to the insurance claims from the Westchester County. There is preliminary evidence that the warnings generated by some of the data sources (survey and phone in particular) lead the clinical sources.

We have also evaluated the Epi-SPIRE anomaly detection mechanisms in a synthetic environment in which site-specific or regional outbreaks are simulated. The results indicate that the pathogen release can be detected within 4 days for acceptable false alarm levels.

6. SUMMARY

In this paper, we describe the architecture and implementation of the Epi-SPIRE prototype, which is a novel health activity monitoring (HAM) system that generates alerts from environmental, behavioral, and public health data sources. A model-based approach is used to develop the disease and behavior models from multi-modal heterogeneous data sources. This system has been successfully validated in a number of scenarios involving infectious disease outbreak.

7. REFERENCES

[1] Glass, GE, T. L. Yates, J. B. Fine, T. M. Shields, J. B. Kendall, A. G. Hope, C. A. Parmenter, C.J. Peters, T. G. Ksiazek, C.-S. Li, J. A. Patz and J. N. Mills. "Satellite imagery characterizes local animal reservoir populations of Sin Nombre virus in the southwestern United States," Proc. National Academy of Science 99:16817-16822. (December 23, 2002)

[2] Glass, G. E. "Public health applications of near real time weather data". 6th Earth Sciences Information Partnership Conf. 2001.

[3] Klein, S. L., A. L. Marson, A. L. Scott, GE Glass, "Sex differences in hantavirus infection are altered by neonatal hormone manipulation in Norway rats," Soc Neuroscience 2001.

[4] Klein, S. L. , A. L. Scott, G. E. Glass, "Sex differences in hantavirus infection: interactions among hormones, genes, and immunity," Am Physiol Soc. 2001.

[5] Glass G. E. "Hantaviruses. Climate Impacts and Integrated Assessment," Energy Modeling Forum 2001.

[6] S. Bhola, R. Strom, S. Bagchi, and Y. Zhao, "Exactly-once Delivery in a Content-based Publish-Subscribe System," Dependeable Systems and Networks 2002.

[7] Y.-C. Chang, L. D. Bergman, V. Castelli, C.-S. Li, M.-L. Lo, and J. R. Smith, ``The Onion Technique: Indexing for Linear Optimization Queries," ACM SIGMOD 2000, May, 2000.

[8] Belle L. Tseng, Ching-Yung Lin, and John R. Smith. Real-Time Video Surveillance for Traffic Monitoring Using Virtual Line Analysis IEEE ICME, Lausanne, Switzerland, August 2002.

Appendix B:An Evaluation of Over-The-Counter Medication Sales for Syndromic Surveillance

Murray Campbell Chung-Sheng Li Charu Aggarwal Milind Naphade
Kun-Lung Wu Tong Zhang

IBM T.J. Watson Research Center
Yorktown Heights, NY 10598
USA

Abstract

Early and reliable detection of disease outbreaks is an important problem for public health. Syndromic surveillance systems use pre-diagnostic data sources to attempt to improve the timeliness of outbreak detection. This paper describes a number of approaches to evaluating the utility of data sources in a syndromic surveillance context. We show that there is some evidence that sales of over-the-counter medications have value for syndromic surveillance.

1 Introduction

Syndromic surveillance refers to the use of pre-diagnostic health-related data for early detection of disease outbreaks. With recent concern over the threat of bioterrorism, as well as the appearance of new disease threats (e.g., SARS), syndromic surveillance is being looked to as a means to improve the timeliness of public health surveillance.

The development of a useful syndromic surveillance system depends in part on the identification of data sources that have value in predicting disease outbreaks. This paper will focus on methods for assessing the value of data sources for predicting disease outbreaks. We will examine a number of different approaches that use retrospective analysis to evaluate data sources.

A frequently cited example of a data source that is presumed to be useful for syndromic surveillance is the sale of over-the-counter (OTC) medications. We will apply our evaluation approaches to a large, multi-year, multi-city data set and show that there is some evidence that OTC medication sales may be useful for syndromic surveillance.

2 Background and Related Work

Syndromic surveillance (also referred to in the literature as early detection of disease outbreaks, pre-diagnosis surveillance, non-traditional surveillance, enhanced surveillance, non-traditional surveillance, and disease early warning systems) has received substantial interest recently, especially after Sept. 11, 2001 [3, 5, 9, 12, 13, 14, 15].

A number of studies have been devoted to investigating various data sources, such as the text and the ICD-9 diagnosis code of the chief complaints from emergency department [1, 2, 6, 11], 911 calls [4], and over-the-counter (OTC) drug sales [8].

There are at least three different classes of approaches to evaluating the utility of a data sources for syndromic surveillance. The first approach is based on the measuring the correlation between a target data source and a gold standard (diagnostic) data source [16]. A second approach is to use the target data source to better predict values in the gold standard data source. A third option is to identify "events" (i.e., disease outbreaks) in a gold standard data source, and assess the timeliness of alarms produced by a detection algorithm operating on the target data source. The tradeoff between timeliness and false alarms can be assessed using the AMOC approach [7].

3 Data

There are two data sets that will be used in this study. The first, which we will call OTC, is a weekly summary of unit sales of upper respiratory over-the-counter medication sales for ten cities (Baltimore/Washington, Charlotte, Chicago, Dallas, Milwaukee, New York, Norfolk, Orlando, Pittsburgh, and Seattle) for a three-year period (2000-2002). The first data point is for the week ending on 1/9/2000, and the last data point is for the week ending 12/29/2002. For each city, sales are reported in eight categories: four types

(Cold, Allergy, Cough, and Sinus), and two target groups for each type (Adult and Pediatric).

The second data set, which we will call CL, consists of anonymized medical insurance claims records. The records are from the same ten cities as for OTC, and cover the same three-year period. Each record consists of a unique (anonymized) patient identifier, a date of service, up to four ICD-9 (diagnosis) codes, and a city name. There are a total of about 22.5 million records. The ICD-9 codes were chosen by the data provider, Surveillance Data, Inc., to be relevant to upper respiratory infections. The number of insurance claims were aggregated by city to weekly totals aligned with the OTC data.

For the purposes of this study, the OTC data set is the target data source, i.e., OTC will be assessed for value in syndromic surveillance. CL is the gold standard data source, as it contains diagnostic information about actual disease.

4 Approaches

4.1 Lead-Lag Correlation Analysis

One approach to evaluating a data source for syndromic surveillance is to conduct a lead-lag correlation analysis on the data source with respect to a gold standard data source. This consists of computing the correlation between the two time series for a range of lead-lag times, and identifying the lead-lag time at which the correlation is maximized. It can be useful to remove trends before analyzing.

Although a correlation analysis can give a global view of the lead time of a target data source, syndromic surveillance is typically more interested in the lead time prior to increasing levels of disease. This suggests an alternative approach where a correlation analysis is performed on a number of shorter time segments that contain the initial stages of disease outbreaks.

In Section 5.1 we will apply this method to the data sets described in Section 3, and assess the value of OTC data for syndromic surveillance.

4.2 Regression Test of Predictive Ability

This section describes another approach to evaluating the usefulness of a target data source by posing it as a prediction problem. More specifically, we are interested in predicting certain quantities associated with the gold standard data source, and want to see whether by including the target data, we are able to make better predictions.

This approach can be generally regarded as time-series forecasting. If we can forecast a quantity A more accurately using a quantity B under a certain metric, then we say that B contains useful information for predicting A.

We now give a general description of this approach. Assume that the quantity of interests is presented sequentially as a time-series

$$\{Y\} = \{\cdots, Y_0, Y_1, \cdots, Y_t, \cdots\}.$$

We want to predict the future values of this time-series based on some side-information (which may includes the historical values of Y we observed so-far), represented as another time-series of vectors:

$$\{\mathbf{X}\} = \{\cdots, \mathbf{X}_0, \mathbf{X}_1, \cdots, \mathbf{X}_t, \cdots\}.$$

Each Y_t is a real-valued number, observed at time t, which we are interested in. Each $\mathbf{X}_t$ is a real-vector, which encodes all of the side information that we hope are useful for predicting the $\{Y\}$ series.

To this end, we assume that at each time t, based on the current side-information X_t, we would like to predict Y_{t+f}, which is the value of the Y series f-steps in the future (where $f > 0$ is an integer). We assume that the predictor $p_f(\mathbf{X}_t)$ has a linear form as

$$Y_{t+f} \approx p_f(\mathbf{X}_t) = \mathbf{w}_f^T \mathbf{X}_t,$$

where $\mathbf{w}_f$ is a weight vector (parameter of our model) that characterizes the predictor p_f. The parameter $\mathbf{w}_f$ can be estimated from the data (as we will describe later).

Given a predictor, represented as a weight vector $\mathbf{w}$, we can measure its quality using a certain figure of merit. In this study, we employ the commonly used least-squares error criterion, defined as

$$R_f(\mathbf{w}, [T_1, T_2]) = \frac{1}{T_2 - T_1 + 1} \sum_{t=T_1}^{T_2} (\mathbf{w}^T \mathbf{X}_t - Y_{t+f})^2.$$

The number $R_f(\mathbf{w}, [T_1, T_2])$ measures in the interval $[T_1, T_2]$, how well we can predict from X the sequence Y f-steps in advance with the weight vector $\mathbf{w}$.

The weight vector can be estimated from the historical data using least-squares regression:

$$\hat{\mathbf{w}}_{f,T} = \arg\min_{\mathbf{w}} \sum_{t=1}^{T-f} (\mathbf{w}^T \mathbf{X}_t - Y_{t+f})^2. \qquad (1)$$

Now assume that we observe the sequences $\mathbf{X}$ and Y, up to some point T. To check how useful is $\mathbf{X}$ for predicting Y, we divide the time period into K consecutive blocks (for simplicity, assume that T is divisible by K): $I_j = [T_j, T_{j+1}]$ for $j = 0, \ldots, K-1$, where $T_j = jT/K$. Now we can use a single number

$$r_f(\mathbf{X}, Y) = \frac{1}{K} \sum_{j=1}^{K-1} R_f(\hat{\mathbf{w}}_{f,T_j}, [T_j, T_{j+1}]) \qquad (2)$$

to measure the usefulness of $\mathbf{X}$ for predicting Y (f-steps in the future). That is, we train a predictor $\hat{\mathbf{w}}_{f,T_j}$ using least squares regression (1) with data observed up to jT_0, and then test on data from jT_0 to T_{j+1}, for $j = 0, \ldots, K-1$, and then average the results. The smaller $r_f(\mathbf{X}, Y)$ is, the more useful $\mathbf{X}$ is for predicting Y. Therefore using (2), we can compare the usefulness of different side informations $\mathbf{X}$ and $\mathbf{X}'$.

In Section 5.2, we compute the corresponding $r_f(\mathbf{X}, Y)$ numbers with and without including the OTC data in the side information $\mathbf{X}$. Our results suggest the usefulness of the OTC data in public health surveillance.

4.3 Detection-Based Approaches

For the detection-based approaches we assume that disease outbreak events are labeled in the gold standard data set, and an outbreak detection algorithm operates on either the the target data set or the gold standard data set. Using the AMOC approach, we are able to assess the lead time provided by the target data source over a range of practical false alarm rates.

4.3.1 Supervised Algorithm for Outbreak detection in OTC data

The supervised outbreak detection algorithm utilized the previously supplied data in order to determine various aspects of the algorithm. The supervised algorithm required a number of components in order to perform the detection:

(1) Determination of features to be used, and the proper way to combine channels.

(2) Creation of streams of anomalies.

(3) Conversion of the anomaly streams into the alarm level using the information from (1).

This supervision was done in two forms:

(1) Feature Selection: Since multiple channels of information were available, which channels provided the greatest level of connection between the channels and actual outbreaks?

(2) Combination of Multiple Channels: How do we combine the signals from multiple channels in order to create one integrated alarm level which was most effective for detecting the outbreak?

In order to perform feature selection, we used the same OTC data set (provided by SDI) as described in the other sections. The first step was to determine which of the channels were most discriminatory for the purpose of distinguishing the biological outbreak from the background noise.

Let us assume that for each site i, the value indicating the channel specific information (absentee behavior, phone calls, pharmacy buying behavior) at time t is denoted by $y(i,t)$. The first step was to convert the data into statistical deviation levels which could be compared across different features. Thus, each stream of data was converted into a statistical stream of numbers indicating the deviation level with respect to the prior window of behavior of width W. The statistical deviation value for a given stream i at time t was denoted by $z(i,t)$. The value of $z(i,t)$ was found by first fitting the prior window of with W with the polynomial function $f(t)$. The deviation value at time $t0$ was then defined as follows:

$$s(i) = \sqrt{\sum_{t=t0-W}^{t0} (f(t) - y(i,t))^2/(W-1)} \qquad (3)$$

The value of W used was based on the last 16 reports. This statistical deviation is also referred to as the z-number. This value provides an idea of how far the stream of data deviates from the normal behavior and gives an intuitive understanding of the level of anomaly at a given tick. Then, the statistical deviation $z(i,t0)$ at time $t0$ is denoted by:

$$z(i,t0) = (f(t0) - y(i,t0))/s(i) \qquad (4)$$

These alarm values could be used in order to determine the value of each channel in the training data. A particular channel was found to be useful when this value was found to be larger than a pre-defined threshold of 1.5. For example, by using this technique we were able to eliminate the allergy channel for the purpose of detection of the flu infections. For example, this behavior was illustrated by the allergy channel in the OTC training data. We have also illustrated the AMOC curve for the allergy channel in the same figure. We note that the AMOC curve for the allergy channel was particularly poor, because it seemed to be uncorrelated to the seasonal outbreaks in the data.

Once these features were selected, they could be used on the test data for computing the statistical deviation values using the same methodology as discussed above. Thus, a separate signal was obtained from stream. The next step was to combine the deviation values from the different sites and channels to create one composite signal. A supervised training process was utilized to determine the optimal functional form for the test data. This was achieved my finding the composition which maximized the area under the AMOC curve.

Once each channel had been converted into a single composite signal, they need to be combined to create a combination signature. For example, let $q1(t)$, $q2(t)$ and $q3(t)$ be the signatures obtained from three different channels. The combination signature was defined as the expression:

$$C(t) = c1 \cdot q1(t) + c2 \cdot q2(t) + c3 \cdot q3(t) \qquad (5)$$

Here $c1$, $c2$ and $c3$ were coefficients which were also determined by minimizing the latency of detection on the training data. As a normalization condition, it is assumed that

the coefficients satisfy the following condition for the constant C':

$$c1^2 + c2^2 + c3^2 = C' \quad (6)$$

It is necessary to use the above condition for scaling purposes. In order to determine the optimal alarm we found values of c_1, c_2, and c_3, which optimized the area under the AMOC curve. This provides the combination signature.

4.3.2 Modified Holt-Winters forecaster

One of the unsupervised detectors used was a modified Holt-Winter forecaster [10]. The forecaster generate a z-value for each tick of a data channel, representing the deviation of observed data from the predicted one. A z-value is computed as follows:

$$z = (\Delta - \mu)/\sigma,$$

where Δ is the difference between observed and predicted data, and μ and σ are the mean and standard deviation, respectively, of these Δ differences in the past.

A Holt-Winters forecaster assumes that a time series, $X_1, \cdots, X_N$, can be modeled in terms of three key components: the average $\overline{X}_N$, the trend T_N and the daily seasonality factors $F_{N-D+1}, \cdots, F_N$, where D is the number of days in the week for which there are observed data. The average is the exponentially smoothed level value of all the time series values. The trend is the exponentially smoothed slope of all the N time series values. The daily seasonality factors are exponentially smoothed values reflecting the deviation from linearity attributable to the different days of the week. The seasonality factors can have either a multiplicative or additive effect. In our implementation, we chose the additive variant. A Holt-Winters forecaster attempts to accurately capture these three key components of a time series. It can deal with special events, such as holidays or special days where data are missing.

4.3.3 Forecasting based on Multi-channel Regression

A simple prediction strategy that can combine single and multi-channel prediction is to set up the problem as a linear regression. As usual, the deviation of the actual value from the predicted value as a measure of abnormality. We set up a system of linear equations as shown below.

Let the observation stream of a single channel from among the multiple OTC sales channels be $[y_1, \ldots, y_M]$. Consider using the past J observations to derive the regression parameters while using the past K samples for actually predicting the $K + 1^{th}$ observation. The number of variables to be estimated from the past J samples is K.

$$\begin{bmatrix} y_{M-1} \\ y_{M-2-1} \\ \cdots \\ y_{M-J-1} \end{bmatrix} = \begin{bmatrix} y_{M-2} & \cdots & y_{M-K-1} \\ y_{M-3} & \cdots & y_{M-K-2} \\ \cdots & \cdots & \cdots \\ y_{M-J-2} & \cdots & y_{M-K-J-1} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \cdots \\ w_K \end{bmatrix} \quad (7)$$

or using matrix notation:

$$Y = A_y W, \quad (8)$$

With this overdetermined system of equations ($J >$K) we then calculate the least squares fit to this as shown in Eq 9:

$$W = (A_y^t A_y)^{-1} A_y^t Y \quad (9)$$

Assuming linear independence among columns of matrix A, $A^t A$ is non singular and the generalized inverse $(A^t A)^{-1}$ exists. We calculate the weight vector W after every update. Thus for each observation y_M we calculate the prediction aW, a being a row vector $[y_{M-1} y_{M-2-1} \ldots y_{M-J-1}]$. If the residual between the actual value and the predicted value is positive we use this difference as a measure of abnormality and probability of an outbreak. Equation 7 can be extended to make the prediction based on multiple channels. For example the matrix A can be created by combining multiple channels. Equation 10 shows past samples from two channels $[y_1, \ldots, y_M]$ and $[x_1, \ldots, x_M]$ being used to predict the current observation of channel Y.

$$Y = [A_y A_x] \begin{bmatrix} W_y \\ W_x \end{bmatrix} \quad (10)$$

Using the above formulation we can predict the current value of sales of any of the OTC channel based on values of sales in the same channel as well as based on values of sales in additional channels.

5 Experiments

5.1 Lead-Lag Correlation Analysis of OTC Data

The lead-lag correlation analysis approach requires us, for each city, to compute the correlations corresponding to various possible lead-lag times. In Figure 5.1, we examine offsets ranging from five weeks prior to five weeks after. The ten solid lines are the correlation values for each of the ten cities. The dashed line is the mean of those values. The peak correlation is between one and two weeks leading, i.e., OTC leading CL by one to two weeks. If a quadratic is fitted to the dashed line, the maximum is at 1.7 weeks.

The provides evidence, albeit somewhat weak, that OTC leads CL and may have value for syndromic surveillance. Clearly there is a wide discrepancy on the correlation between OTC and CL across the different cities, and this needs further investigation.

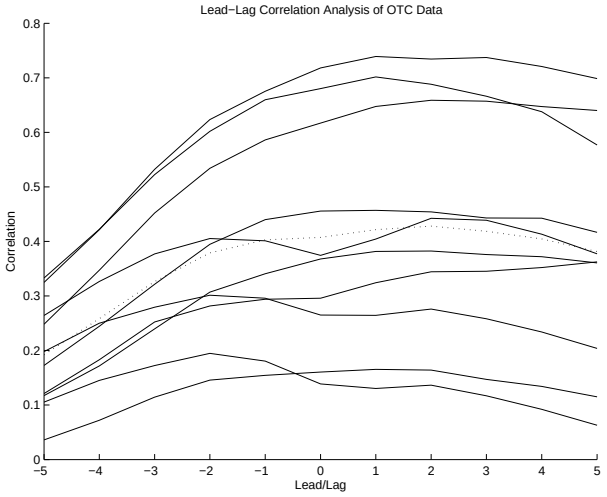

Figure 1. Lead-Lag correlation analysis experiment

5.2 Regression Test of the Predicative Value of OTC

We study the usefulness of OTC for predicting insurance claims using the approach described in Section 4.2. Since the OTC data are weekly based, we shall form the time series on a weekly basis. In particular, we convert the insurance data into weekly data aligned with the OTC data.

In this experiment, we consider different cities separately. That is, we do not consider possible inter-city correlations. For each city, we let OTC_t be the total number of OTC sales in week t, and CL_t be the number of insurance claims in week t. Since in public health surveillance, we are mostly interested in sudden outbreaks of diseases, we are interested in the log-ratio of the number of insurance claims in consecutive weeks. That is, at week t, the Y variable is given by

$$Y_t = \log_2(CL_t/CL_{t-1}).$$

One may also use other quantities, such as whether the insurance claims next week is higher than this week by a certain amount (or whether Y_t is larger than some threshold).

We consider a few possible side information $\mathbf{X}$, which we list below.

- $\mathbf{X}^1$: Using constant side information: $\mathbf{X}^1_t \equiv [1]$. This leads to a predicator that predict Y_t using its historic mean.
- $\mathbf{X}^{CL}$: In addition to the above, we also include historical observations of the insurance claim data itself (the log ratio of the current number of claims over the claims of the previous week) as side-information: $\mathbf{X}^{CL}_t = [Y_t, 1]$.
- $\mathbf{X}^{OTC}_t$: We include the constant one and the OTC data into the side-information:

$$\mathbf{X}^{OTC}_t = [\log_2(OTC_t/OTC_{t-1}), 1].$$

- $\mathbf{X}^{CL-OTC}_t$: We include all of the above quantities into the side-information:

$$\mathbf{X}^{CL-OTC}_t = [\log_2(OTC_t/OTC_{t-1}), Y_t, 1].$$

Since this framework is quite flexible, various other configurations can also be studied. For our purpose, we are able to make interesting observations from this particular configuration. Variations will lead to similar results.

Applying the notation in Section 4.2, for each city, we divide the time series into $K = 20$ blocks, and compute the $r_f(\mathbf{X}, Y)$ number in (2) for $f = 1, 2$ and each side information listed above. We then average the results over the ten cities, and report the averaged numbers in Table 1. From the table, we can see that the OTC data has a small predicative power for the insurance claims data CL. One may also do an experiment in the reverse order (that is, use historical CL data to predict the future OTC sales). In this case, for $f = 1$, the predictive performance for OTC sales, measured by the r_f value, degrades from 0.0217 (without CL in the side-information) to 0.0221 (with historical CL data in the side-information). Therefore these experiments provide some evidence suggesting that OTC changes precede CL changes.

	$\mathbf{X}^1$	$\mathbf{X}^{CL}$	$\mathbf{X}^{OTC}$	$\mathbf{X}^{CL-OTC}$
$f = 1$	0.0287	0.0265	0.0285	0.0261
$f = 2$	0.0287	0.0291	0.0280	0.0287

Table 1. Averaged $r_f(\mathbf{X}, Y)$ numbers over ten cities

Although effects shown in Table 1 are relatively small, we believe they are still indicative statistically. Since we average our results over ten cities, we may also check the variation over different cities. In particular, in seven out of ten cities, $r_1(\mathbf{X}^{OTC}, Y)$ is smaller than $r_1(\mathbf{X}^1, Y)$; also in seven out of ten cities, $r_2(\mathbf{X}^{OTC}, Y)$ is smaller than $r_2(\mathbf{X}^1, Y)$. This comparison is consistent with results in Table 1, and justifies from a slightly different point of view that statistically, the OTC data is (weakly) useful for predicting future insurance claims.

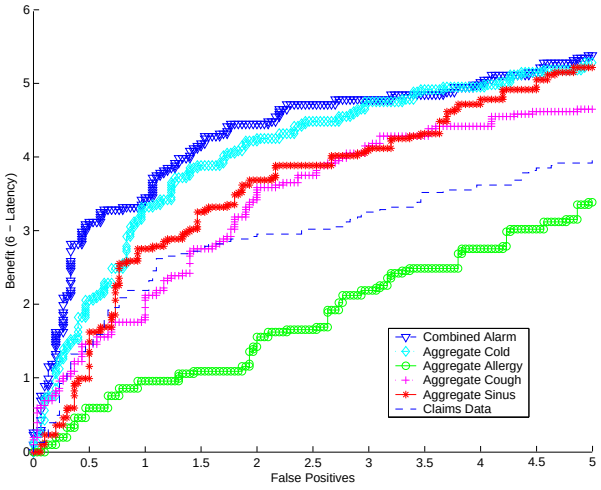

Figure 2. The AMOC curves generated by the Supervised method illustrate that various OTC categories are more timely than claims.

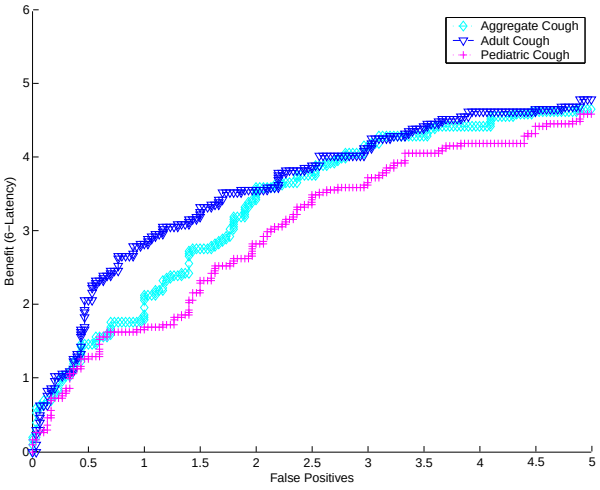

Figure 3. The AMOC curves generated by the Supervised Method illustrate that there is differentiation between Adult and Pediatric cough medication sales.

5.3 Results From Detection-Based Approaches

5.3.1 Supervised Method

Once the features have been selected, and the proper way for construction of the combination signature was determined, the actual alarm level construction on the data was straightforward. The deviation values for the data were computed in an exactly identical way to the training data, and the combination was created to output the corresponding alarm levels at each tick. In Figure 5.3.1, we have illustrated the behavior of the detection algorithms. Once interesting observation was that the OTC data was always more effective than the claims data. In fact, in most cases, the OTC data acted as a "leading indicator" over the claims data. It is also interesting to note that the adult and pediatric data illustrated differential behavior in terms of the speed and quality of the detection. An example of this is illustrated in Figure 5.3.1.

5.3.2 Modified Holt-Winters forecaster

Even though the OTC data were weekly data, the detector treated them as daily data and assumed that there were 3 days in a week. It used the past 6 OTC data points to predict the next OTC sale.

While there was some variability across different categories of OTC medication sales, over a wide range of false alarm rates the Holt-Winters forecaster showed a two week lead time for OTC over Claims. Sinus medication sales were observed to have the best lead times overall.

5.3.3 Forecasting based on Multi-channel Regression

Using the OTC data we experimented with different values of J and K (see Section 4.3.3 for single and multichannel prediction based outbreak detection. Based on our experiments we found that sales of adult drugs were more informative about the outbreakss and had a lead time of between 2 and 3 weeks over claims. We also found encouraging empirical evidence that the use of multiple channels resulted in a better lead time for predicting outbreaks over single channel prediction. Figure 4 shows the AMOC curve using the adult cold channel for predicting outbreaks. It also shows the benefit of using adult cold and adult cough to predict adult cold sales and use the deviation to detect outbreaks although this benefit is evident only for small values of false alarms as seen in the AMOC curve

6 Conclusions and Future Work

We have shown a number of different approaches to assessing the value of a data source for syndromic surveillance, and evaluated over-the-counter medication sales using these approaches. The appears to be evidence from each of these approaches that OTC medication sales are a leading indicator for disease outbreaks.

There are a number of limitations in this study. The data sets were aggregated weekly, which reduces the precision regarding estimates of the timeliness of OTC. This type of study should be repeated with daily data. The detection-based experiments identified only those outbreaks that occurred at the beginning of the seasonal rise in respiratory disease. A more careful study could examine finer grain

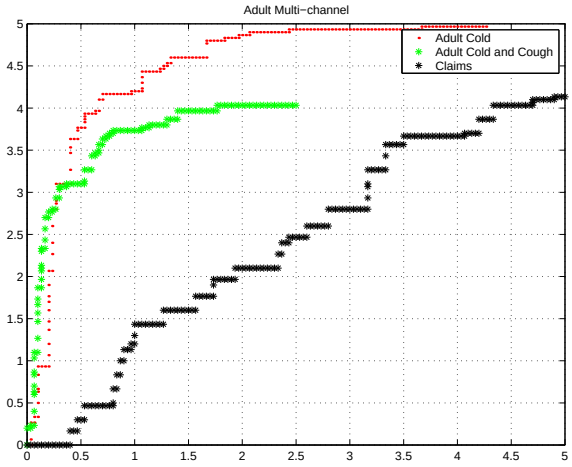

Figure 4. The Adult Cold sales were found to be the best indicator for the outbreaks with $J = 15, K = 2$ and $J = 20, K = 1$ respectively for single channel and multi-channel prediction. The Adult Cold and Cough sales were used in the two channel prediction.

disease outbreaks, preferably those that have been studied and verified by public health. This study was retrospective, looking only at historical data. A prospective study, using the target data source to predict disease outbreak in real time, would provide greater confidence in the conclusions in this paper.

7 Acknowledgments

We would like to thank Andrew Kress of Surveillance Data, Inc. for supplying the data used in this study. This work is supported by the Air Force Research Laboratory (AFRL)/Defense Advanced Research Projects Agency (DARPA) under AFRL Contract No. F30602-01-C-0184. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the AFRL or DARPA.

References

[1] E. M. Begier, D. Sockwell, L. M. Branch, J. O. Davies-Cole, L. H. Jones, L. Edwards, J. A. Casani, and D. Blythe. The 1979. national capitol region's emergency department syndromic surveillance system: do chief complaint and discharge diagnosis yield different results? *Emerging Infectious Disease*, 9(3):393–396, Mar. 2003.

[2] A. J. Beitel, K. L. Olson, B. Y. Reis, and K. D. Mandl. Use of emergency department chief complaint and diagnostic codes for identifying repiratory illness in a pediatric population. *Pediatric Emergency Care*, 20(6):355–360, Jun. 2004.

[3] J. W. Buehler, R. L. Berkelman, D. M. Hartley, and C. J. Peters. Syndromic surveillance and bioterrorism-related epidemics. *Emerging Infectious Diseases*, 9(10):1197–1204, Oct. 2003.

[4] M. R. Dockrey, L. J. Trigg, and W. B. Lober. An information systems for 911 dispatch monitoring system and analysis. In *Proceeding of the AMIA Sympoisum*, page 1008, 2002.

[5] J. S. Duchin. Epidemiological response to syndromic surveillance signals. *Journal of Urban Health*, 80(2):i115–i116, 2003.

[6] J. U. Espino and M. M. Wagner. Accuracy of icd-9-coded chief complaints and diagnoses for the detection of acute respiratory illness. In *Proceedings AMIA Symposium*, pages 164–168, 2001.

[7] T. Fawcett and F. Provost. Activity monitoring: Noticing interesting changes in behavior. In Chaudhuri and Madigan, editors, *Proceedings on the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 53–62, San Diego, CA, 1999.

[8] A. Goldenberg, G. Shmueli, R. A. Caruana, and S. E. Fienberg. Early statistical detection of anthrax outbreaks by tracking over-the-counter medication sales. *Proceedings of the National Academy of Sciences of the United States of America*, 99(8):5237–5240, Apr. 2002.

[9] T. Goodwin and E. Noji. Syndromic surveillance. *European Journal of Emergency Medicine*, 11(1):1–2, Feb. 2004.

[10] C. Granger and P. Newbold. *Forecasting Economic Time Series*. Academic Press, 1977.

[11] J. Greenko, F. Mostashari, A. Fine, and M. Layton. Clinical evaluation of the emergency medical services (ems) ambulance dispatch-based syndromic surveillance system, new york city. *Journal of Urban Health*, 80(2):i50–i56, 2003.

[12] F. Mostashari and J. Hartman. Syndromic surveillance: A local perspective. *Journal of Urban Health*, 80(2):i1–i7, 2003.

[13] J. A. Pavlin. Investigation of disease outbreaks detected by syndromic surveillance systems. *Journal of Urban Health*, 80(2):i107–i114, 2003.

[14] D. M. Sosin. Syndromic surveillance: The case for skillful investment. *Biosecurity and Bioterrorism: Biodefense Strategy, Practice, and Science*, 1(4):247–253, 2003.

[15] M. M. Wagner, F. C. Tsui, J. U. Espino, V. M. Dato, D. F. Sittig, R. A. Caruana, L. F. McGinnis, D. W. Deerfi eld, M. J. Druzdzel, and D. B. Fridsma. The emerging science of very early detection of disease outbreaks. *Journal of Public Health Management Practice*, 7(6):51–59, Nov. 2001.

[16] R. Welliver, J. Cherry, K. Boyer, J. Deseda-Tous, P. Krause, J. Dudley, R. Murray, W. Wingert, J. Champion, and G. Freeman. Sales of nonprescription cold remedies: a unique method of influenza surveillance. *Pediatric Research*, 13:1015–1017,